



건강음료

판매량 예측을 위한  
최적의 모델과  
분석방법 탐색

박진원 임원기

# 분석 과정

분석 도구



# 데이터 탐색

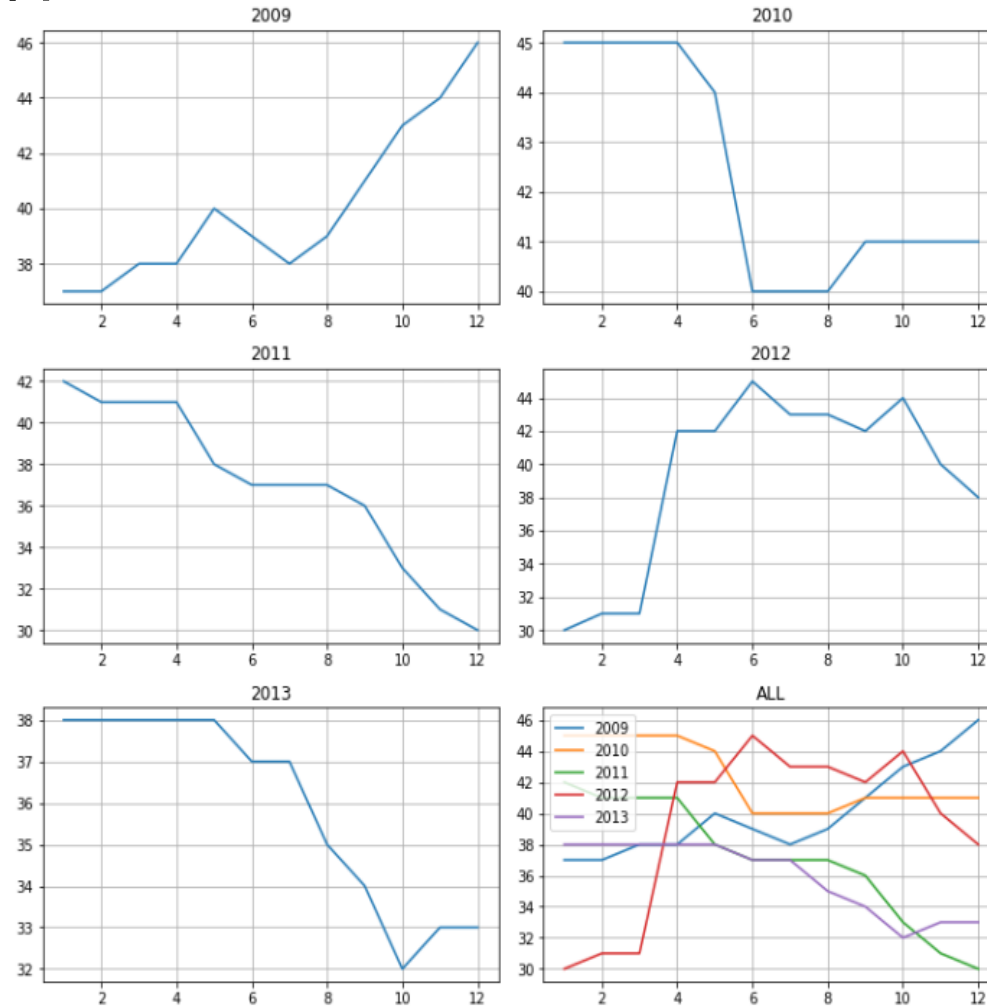
## 변수 파악

	A	B	C	D	E	F	G	H	I
1	YM	CATEGORI	ITEM_CNT	QTY	PRICE	MAXTEMP	SALEDAY	RAIN_DAY	HOLIDAY
2	200901	건강음료	37	1410	1543	4	126753	151	12
3	200902	건강음료	37	1209	1543	9	113399	3014	8
4	200903	건강음료	38	1348	1547	12	129162	1383	9
5	200904	건강음료	38	1377	1500	19	126277	3564	8
6	200905	건강음료	40	1406	1528	25	129584	9288	12
7	200906	건강음료	39	1343	1623	27	123218	9998	8
8	200907	건강음료	38	1313	1600	28	131083	17405	8
9	200908	건강음료	39	1448	1577	29	130040	7704	10
10	200909	건강음료	41	1531	1559	27	131989	4096	8
11	200910	건강음료	43	1670	1586	22	136095	2824	10
12	200911	건강음료	44	1572	1573	12	136152	3684	

- YM : 판매년월
- CATEGORI : 음료 카테고리
- ITEM\_CNT : 상품 품목수
- QTY : 판매량
- PRICE : 가격
- MAXTEMP : 기온
- SALEDAY : 영업(판매)일수
- RAIN\_DAY : 강우일수
- HOLIDAY : 휴일일수

# 데이터 탐색

## 연도별 상품 품목수



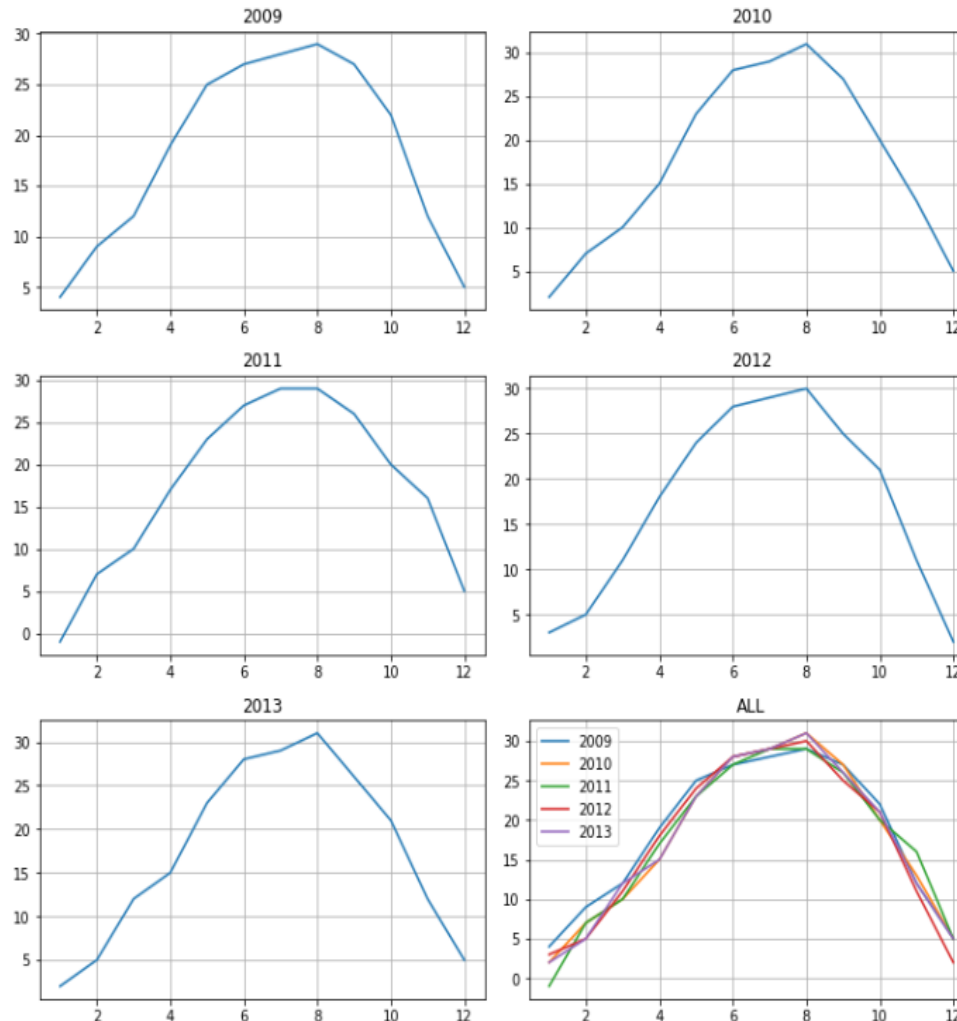
# 데이터 탐색

## 연도별 상품 가격



# 데이터 탐색

## 연도별 기온



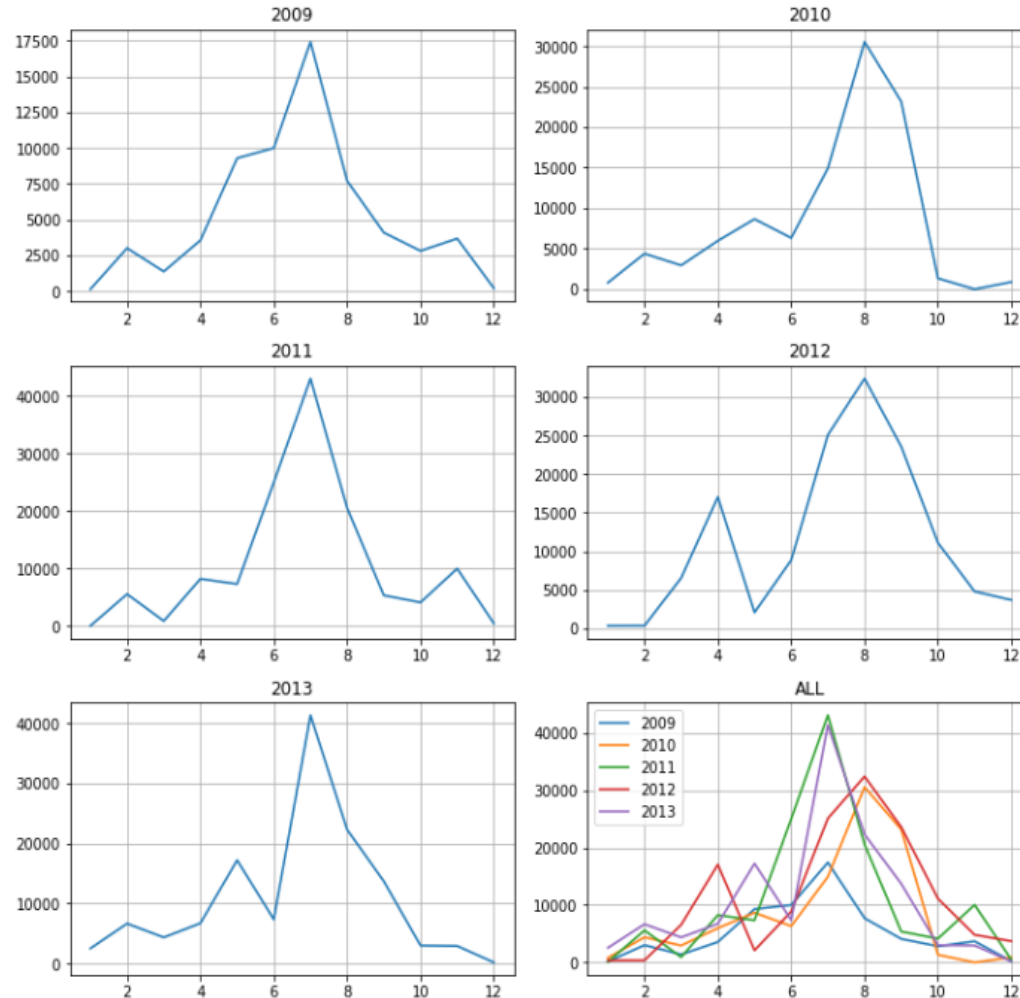
# 데이터 탐색

## 연도별 판매일수



# 데이터 탐색

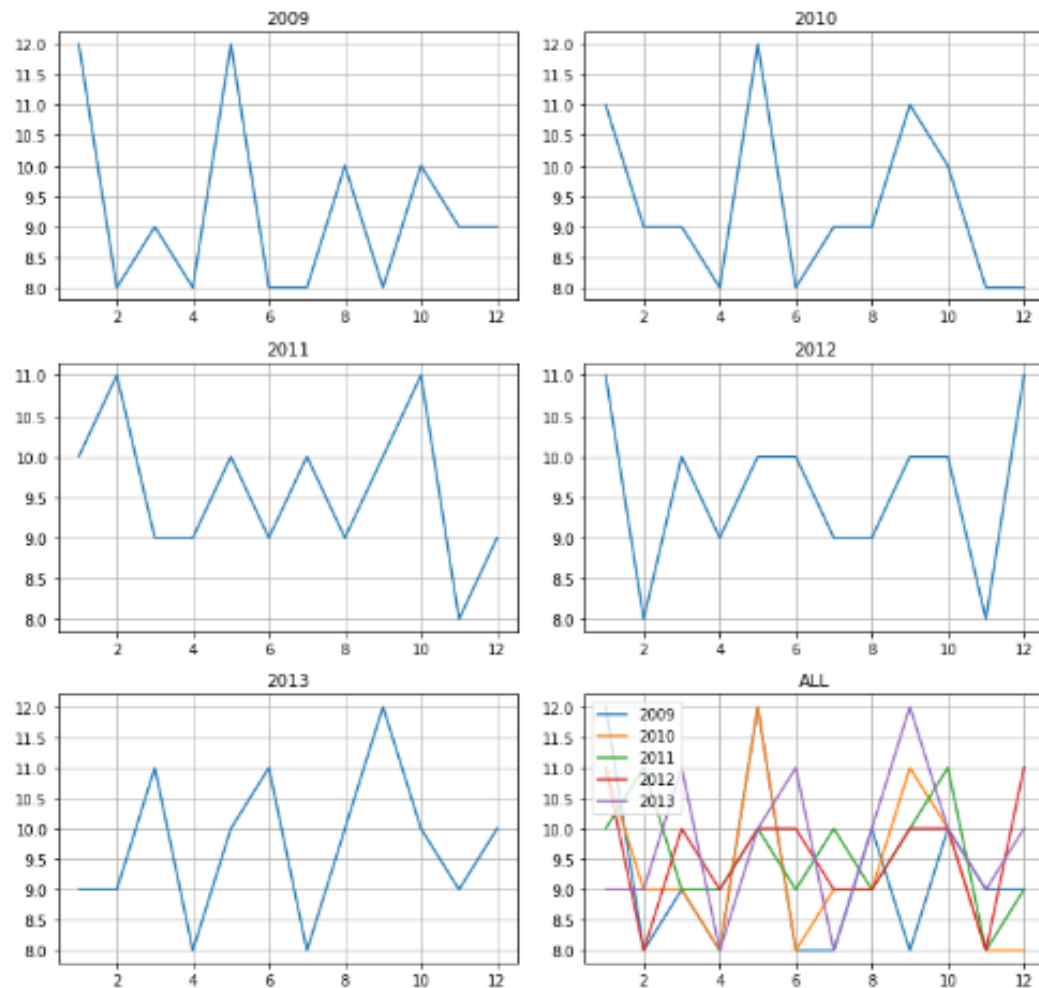
## 연도별 강우일수





# 데이터 탐색

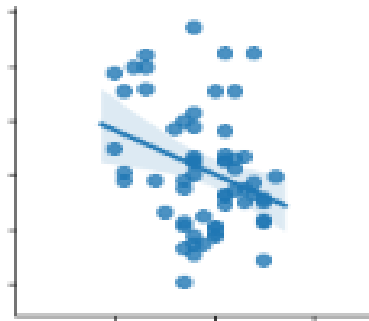
## 연도별 휴일일수



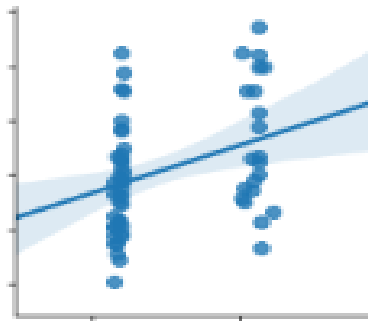
# 데이터 탐색

독립변수와 종속변수 간의 관계

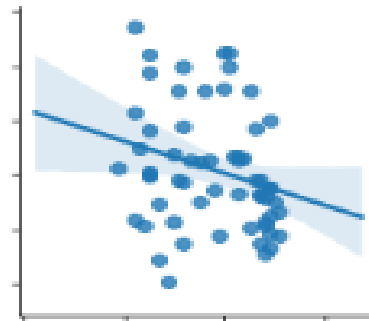
판매량



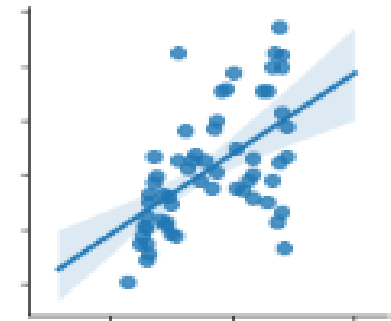
품목수



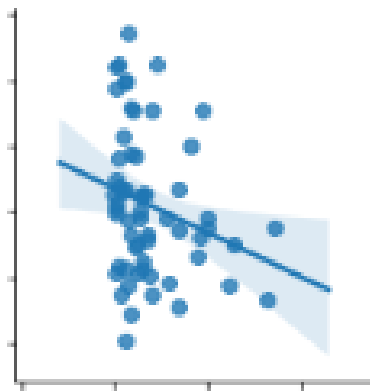
가격



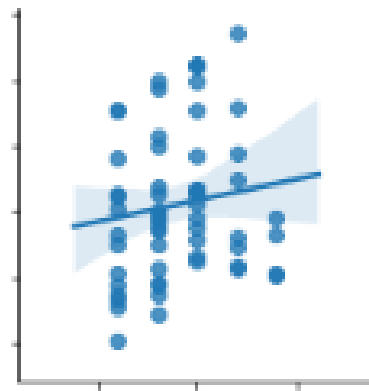
기온



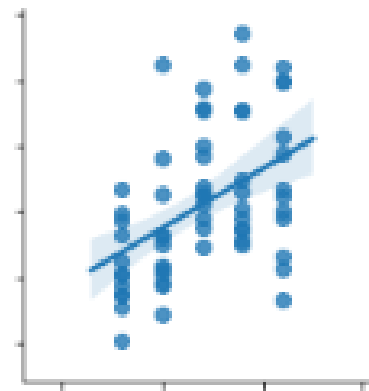
영업일수



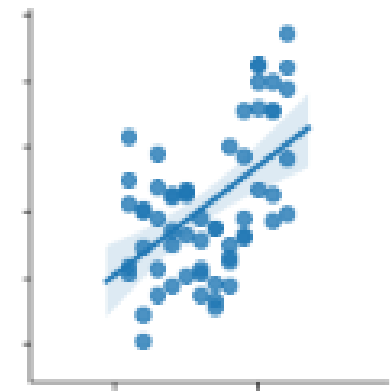
강우일수



휴일일수



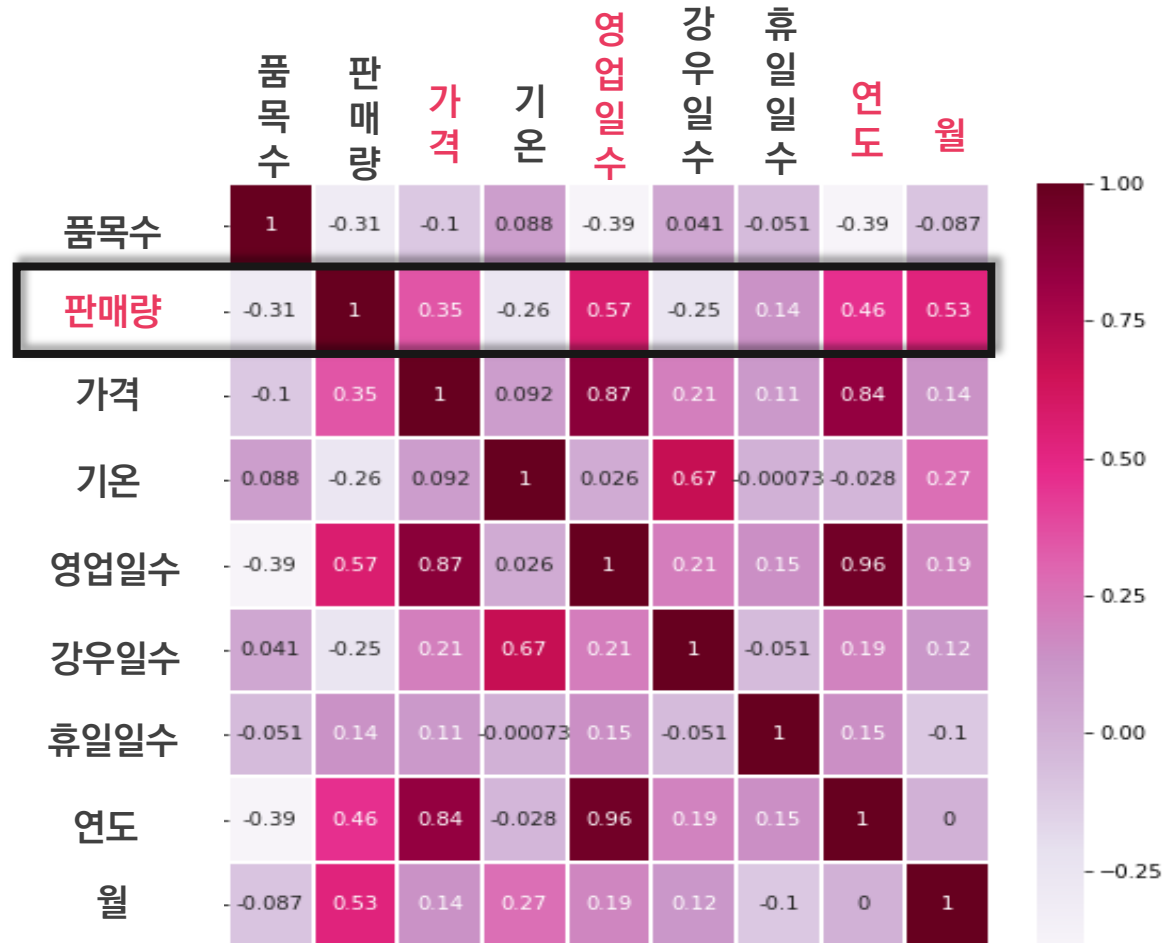
연도



월

# 데이터 탐색

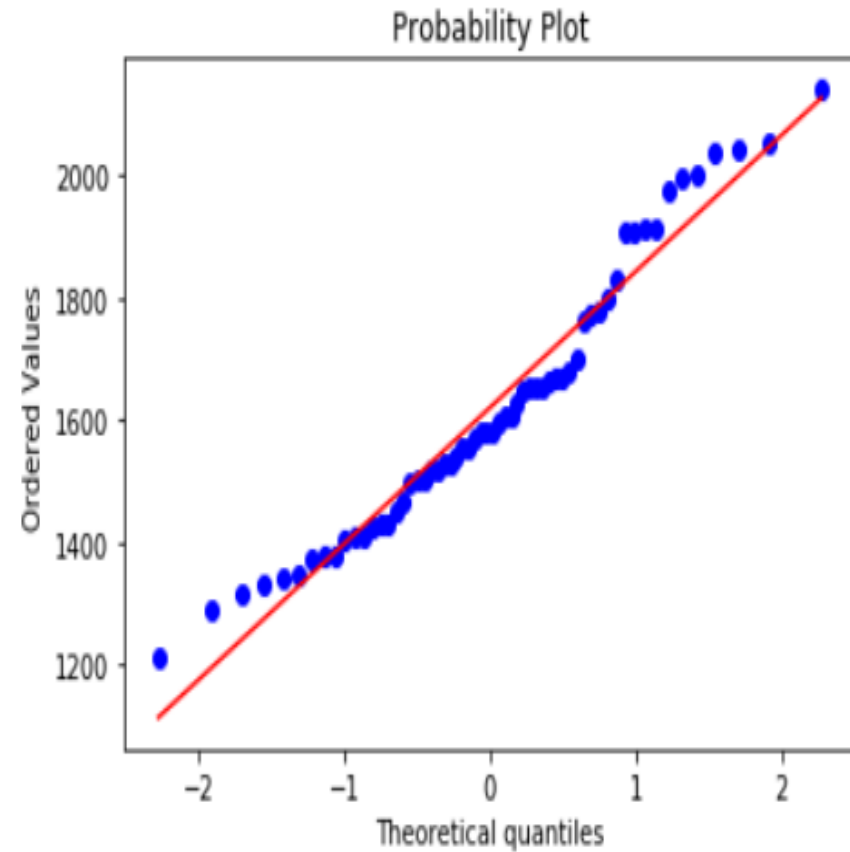
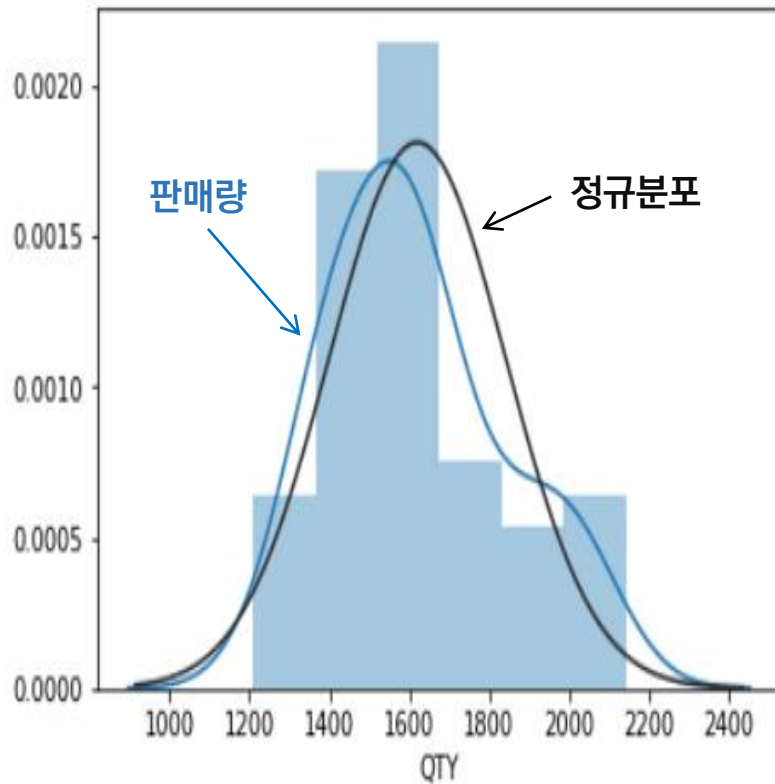
## 판매량과의 상관관계



# 데이터 탐색

종속변수(판매량)

정규성 검정



P-value : 0.0347

유의수준 0.01에서 종속변수가 정규성을 만족한다고 판단

# 모델 평가

## 첫번째 모델

QTY ~ ITEM\_CNT + PRICE + MAXTEMP + SALEDAY + RAIN\_DAY + HOLIDAY

판매량 ~ 품목수 + 가격 + 기온 + 판매일수 + 강수일수 + 휴일일수

	Regression	RMSE	R2	Adj_R2
1	LinearRegression	143.4655	0.3722	-0.0987
2	DecisionTreeRegressor	220.5543	-0.4838	-1.5966
3	SVR	181.0748	-0.0001	-0.7502
4	RandomForestRegressor	162.1092	0.1984	-0.4028
5	ArtificialNeuralNetwork	330.6585	-2.3350	-4.8362

# 모델 평가

## 두번째 모델

QTY ~ ITEM\_CNT + PRICE + MAXTEMP + SALEDAY + RAIN\_DAY + HOLIDAY + MON

판매량 ~ 품목수 + 가격 + 기온 + 판매일수 + 강수일수 + 휴일일수 + 월

	Regression	RMSE	R2	Adj_R2
1	LinearRegression	96.6141	0.7153	0.4306
2	DecisionTreeRegressor	127.7573	0.5021	0.0043
3	SVR	181.0748	-0.0001	-1.0002
4	RandomForestRegressor	53.3110	0.9133	0.8266
5	ArtificialNeuralNetwork	204.4915	-0.2755	-1.5510

# 모델 평가

## 세번째 모델

QTY ~ ITEM\_CNT + MAXTEMP + SALEDAY + RAIN\_DAY + HOLIDAY + MON (-PRICE)

판매량 ~ 품목수 + 기온 + 판매일수 + 강수일수 + 휴일일수 + 월

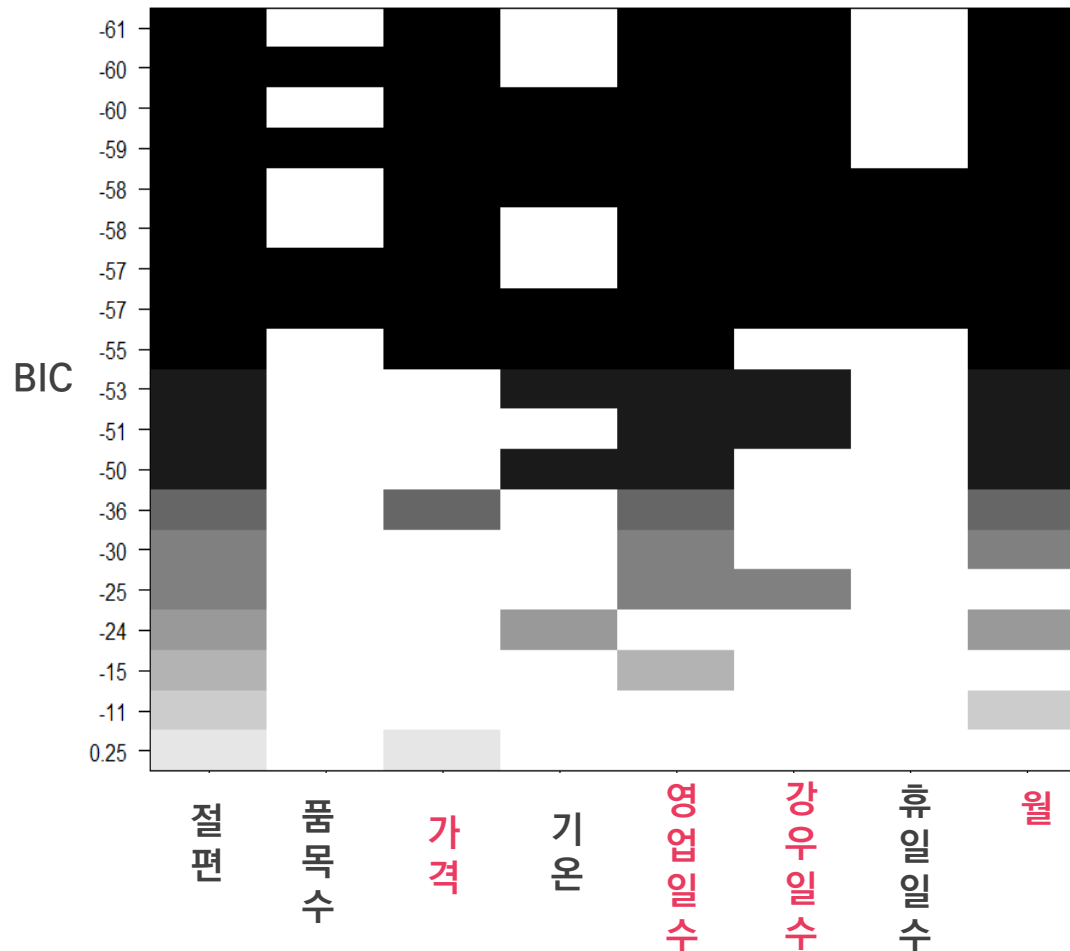
	Regression	RMSE	R2	Adj_R2
1	LinearRegression	100.8029	0.6901	0.4576
2	DecisionTreeRegressor	131.4040	0.4733	0.0783
3	SVR	181.0748	-0.0001	-0.7502
4	RandomForestRegressor	105.2482	0.6621	0.4087
5	ArtificialNeuralNetwork	322.9863	-2.1820	-4.5685

# 변수 선택 with R

BIC : 가격 + 판매일수 + 강우일수 + 월

CP : 품목수 + 가격 + 기온 + 판매일수 + 강우일수 + 월

Adj R2: 모든 변수





# 모델 평가

## 네번째 모델

QTY ~ PRICE + SALEDAY + RAIN\_DAY + MON (-ITEM\_CNT, MAXTEMP, HOLIDAY)

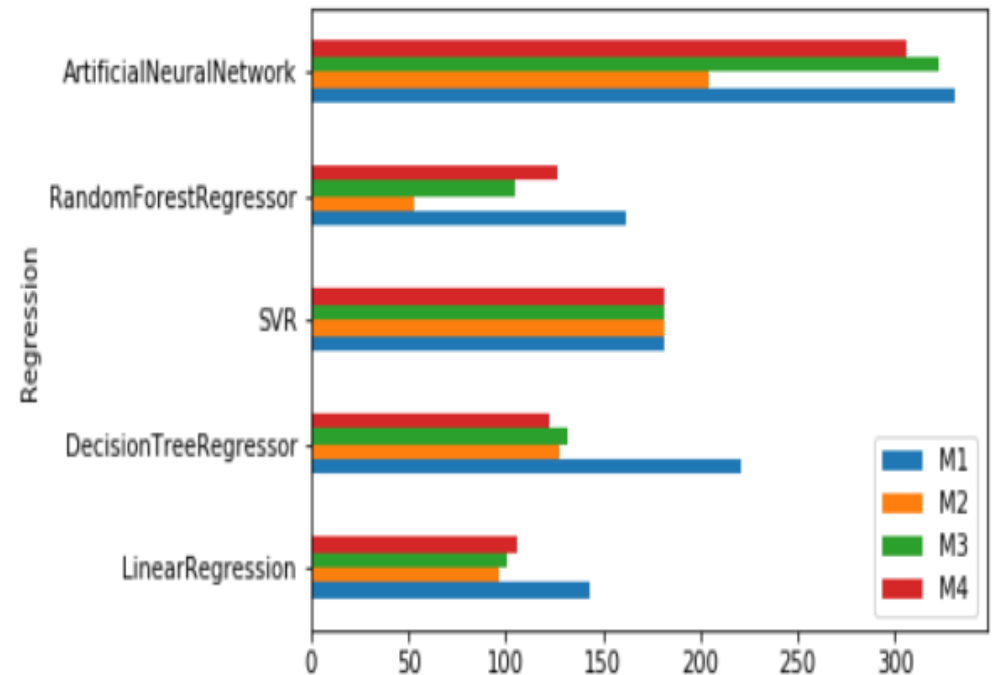
판매량 ~ 가격 + 판매일수 + 강수일수 + 월

	Regression	RMSE	R2	Adj_R2
1	LinearRegression	105.9867	0.6574	0.5203
2	DecisionTreeRegressor	122.8376	0.5397	0.3556
3	SVR	181.0748	-0.0001	-0.4002
4	RandomForestRegressor	126.2652	0.5137	0.3192
5	ArtificialNeuralNetwork	305.7147	-1.8508	-2.9911

# 모델 비교

평균제곱근오차(RMSE)

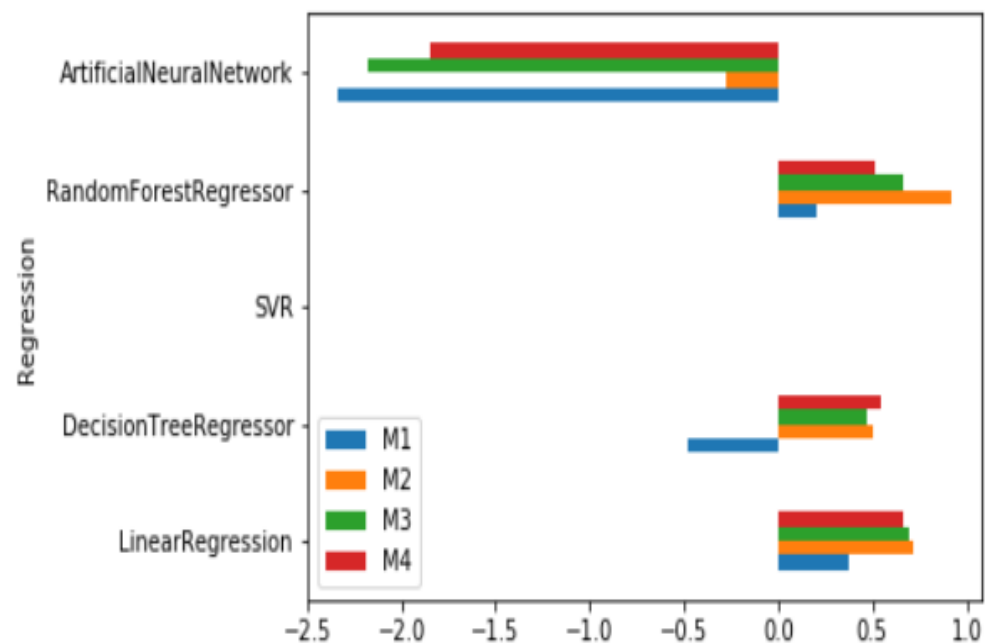
	RMSE_1	RMSE_2	RMSE_3	RMSE_4
Regression				
LinearRegression	143.4655	96.6141	100.8029	105.9867
DecisionTreeRegressor	220.5543	127.7573	131.4040	122.8376
SVR	181.0748	181.0748	181.0748	181.0748
RandomForestRegressor	162.1092	53.3110	105.2482	126.2652
ArtificialNeuralNetwork	330.6585	204.4915	322.9863	305.7147



# 모델 비교

결정계수( $R^2$ )

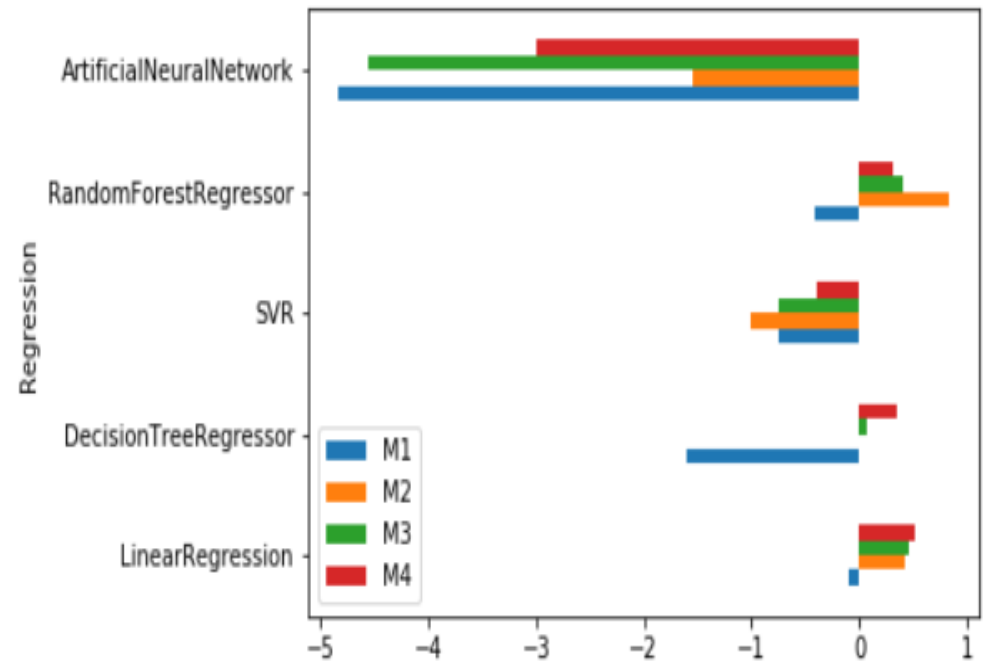
Regression	R2_1	R2_2	R2_3	R2_4
LinearRegression	0.3722	0.7153	0.6901	0.6574
DecisionTreeRegressor	-0.4838	0.5021	0.4733	0.5397
SVR	-0.0001	-0.0001	-0.0001	-0.0001
RandomForestRegressor	0.1984	0.9133	0.6621	0.5137
ArtificialNeuralNetwork	-2.3350	-0.2755	-2.1820	-1.8508



# 모델 비교

수정된 결정계수( $adj R^2$ )

Regression	Adj_R2_1	Adj_R2_2	Adj_R2_3	Adj_R2_4
LinearRegression	-0.0987	0.4306	0.4576	0.5203
DecisionTreeRegressor	-1.5966	0.0043	0.0783	0.3556
SVR	-0.7502	-1.0002	-0.7502	-0.4002
RandomForestRegressor	-0.4028	0.8266	0.4087	0.3192
ArtificialNeuralNetwork	-4.8362	-1.5510	-4.5685	-2.9911



# I 결론

- 건강음료 판매량 예측을 위한 최적의 모델은 **두 번째** 모델
- $QTY \sim \text{ITEM\_CNT} + \text{PRICE} + \text{MAXTEMP} + \text{SALEDAY} + \text{RAIN\_DAY} + \text{HOLIDAY} + \text{MON}$
- 모든 모델에서 회귀분석이 양호한 성능을 발휘
- 하지만 가장 좋은 예측 성능은 **랜덤포레스트** 방법

# Q & A

분석과정과 모든 코드는 구글 드라이브를 참고해주세요

[ PJT - 1조 - mini\_project\_01(py).html ]